

Digital biology: an emerging and promising discipline

Richard W. Morris, Carol A. Bean, Gregory K. Farber, Daniel Gallahan, Eric Jakobsson, Yuan Liu, Peter M. Lyster, Grace C.Y. Peng, Fred S. Roberts, Michael Twery, John Whitmarsh and Karen Skinner^{*,†}

This article examines the role of computation and quantitative methods in modern biomedical research to identify emerging scientific, technical, policy and organizational trends. It identifies common concerns and practices in the emerging community of computationally-oriented bio-scientists by reviewing a national symposium, Digital Biology: the Emerging Paradigm, held at the National Institutes of Health in Bethesda, Maryland, November 6th and 7th 2003. This meeting showed how biomedical computing promises scientific breakthroughs that will yield significant health benefits. Three key areas that define the emerging discipline of digital biology are: scientific data integration, multi-scale modeling and networked science. Each area faces unique technical challenges and information policy issues that must be addressed as the field matures. Here we summarize the emergent challenges and offer suggestions to academia, industry and government on how best to expand the role of computation in their scientific activities.

Introduction

A large, diverse group of scientists gathered at the National Institutes of Health in Bethesda, Maryland last year at a national symposium – Digital Biology: the Emerging Paradigm (<http://www.bisti.nih.gov/2003meeting/archive/agenda.html>). Attendees reported on how computers and the technology-based processes they support are transforming biomedical research. Their presentations and deliberations revealed that today, more than ever before, biomedical scientists are challenged to adopt advanced quantitative and computational methods. Computers are enabling researchers to improve data quality and laboratory efficiency, extend their ability to probe and model

complex biological phenomena and enact or adjust to fundamental changes in the conduct of science. This broad-based ‘quickening’ of discovery driven by computers has the potential to increase scientific breakthroughs and health benefits from biomedical research.

At the meeting, biomedical computing was portrayed as an emerging discipline with well-articulated interests and promising future directions. **Box 1:** The promises of biomedical computing offers some details on the potential of digital biology. The defining interests and challenges of digital biology can be grouped into three areas: (i) scientific data integration; (ii) multi-scale biological modeling and (iii) the networking of science. The symposium depicted biomedical research as approaching the point at which it is essential for scientists, administrators and technologists to understand the promise of computational biology, acknowledge the unique challenges faced by those using computers in biomedical research, accelerate the diffusion of best practices and integrate the views of computational biologists extensively into future programs and plans.

Data integration

Scientific data of interest to biomedical researchers are becoming more complex, heterogeneous and voluminous. Bottlenecks in the use of this data occur owing to our limited capacity to control quality and integrate data from myriad sources, to share data across multiple tasks and to exchange data among different people and organizations. These bottlenecks threaten productivity and efficiency in research activities. Problems with data integration affect all data tasks, including semantic interpretation, data representation, modeling, data storage and query and transaction processing. Despite increasing interest in merging data across experiments or disciplines, it remains difficult to gain unencumbered access to datasets that might be usefully merged. Ultimately the pace of discovery and the health of the biomedical research enterprise will depend on our ability to resolve these issues and thereby extend our ability to locate, access and use disparate data sources.

There is a pressing need for structuring biological knowledge to enable the integration of data and databases across domains, modalities and scales. Part of the solution involves standards for data content and formats. Data integration must accommodate a variety of data types ranging from highly structured data, to complex images, to textual data that are linguistically complex and

Corresponding author: Morris, R.W. (rmorris@niaid.nih.gov).

* Affiliations: Co-authors Morris, Bean, Farber, Gallahan, Jakobsson, Liu, Lyster, Peng, Twery, and Whitmarsh are affiliated with the National Institutes of Health in Bethesda, Maryland. Roberts is in the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS), Rutgers.

† As representatives of various institutes at the National Institutes of Health (NIH), all co-authors actively participated in writing this article. They organized countless discussions and workshops and crafted the agenda and reported findings from a national symposium on digital biology. Most have assisted with the Roadmap initiative of the National Institutes of Health for which biomedical computing is a central thrust. The sum of these efforts made this Opinion article possible. Detailed information about several National Centers of Biomedical Computing, the NIH Roadmap and the NIH Biomedical Information Science and Technology Initiative can be found at <http://www.bisti.nih.gov/ncbc/>.

Available online 26 January 2005

Box 1. The promises of biomedical computing

Promise. Biomedical computing will permit scientists to extract biologically meaningful information from datasets of ever-increasing size, heterogeneity and complexity.

Example. Today PhysioNet [4] provides an open-source platform to generate insights by bridging molecular- and organ-level observations to decipher large collections of recorded physiologic signals. **Promise.** Computational biology will support analysis of genetic variation in disease and drug response so clinicians can predict individual disease risk and tailor therapies.

Example. The HapMap [5] makes disease-gene mapping practical by reducing the genetic variants to be examined in a whole genome scan from 10 million to 500 000.

Promise. Computational biology data and applications will be scaled so that valuable resources and data are made available to the broadest possible scientific community.

Example. Massive amounts of data are generated in high-throughput fashion today that are most useful if broadly shared. Under various Grid and cyber-infrastructure initiatives new linking and translational middleware is the critical enabling technology that makes geographically dispersed resources more widely available [6]. **Promise.** New relationships among biological and quantitative scientists will emerge which benefit both disciplines.

Example. Mathematicians and computer scientists are finding new methods for analyzing data and modeling biological systems. Their findings point to new biological processes that reveal new architectures and performance enhancements for computers.

Promise. Computers will change the conduct of clinical research.

Example. A newly launched initiative at the U.S. National Institutes of Health – National Electronic Clinical Trials and Research (NECTAR) – is intended to streamline clinical research and to accelerate the pace of discovery and application of clinical findings [7].

Promise. New organizational structures will emerge.

Example. Unprecedented inter-governmental partnerships are being created to explore new domains of enquiry and push theory in uncharted directions. One initiative to improve methods in multi-scale modeling, for example, [8] involves all major science agencies of the US Federal government, NIH, National Science Foundation (NSF), National Aeronautics and Space Administration (NASA), and Department of Energy (DoE).

nuanced. Data must be accommodated at various levels of abstraction – molecular, cellular, pathway, organ and even whole organism – and at differing stages of analysis. Data integration must also consider the needs of many biologists who rely on public repositories for the analysis of diverse or novel datasets. Even using highly organized publicly available biological databases, it is often difficult to integrate data.

What are the challenges with integrating scientific data? One is a lack of adequate structured vocabularies and ontologies (computer readable vocabularies, taxonomies and indexes that constitute the central concepts of a scientific discipline) to provide a common basis for describing content and related data. Another stems from the absence of common formats or inter-convertible formats that describe the data. Still another set of challenges is posed by the heterogeneous, context-dependent and extremely varied nature of biological data. Dealing effectively with context requires multiple inter-dependent solutions. Although some scientific disciplines have adopted data standards to allow data integration, most still lack common terminologies, ontologies and repositories. Finally, new algorithms such as Bayesian analysis must be developed that combine heterogeneous

data and form unified, possibly simplified, understanding and knowledge of scientific experiments and behavior.

Biological modeling

Modeling biological systems is another of the major challenges facing biomedical computing. Although some biological processes are well characterized within a narrow time and spatial range, biology is not well understood at the intersections between scales. Most biological models are not scalable, for example models of synaptic connections between neurons do not scale to macroscopic models of the brain. The goal of multi-scale modeling is to provide the scientific community with rigorous, widely recognized methods and descriptive forms to reveal the essential elements needed to simulate complex biological processes across scales.

Similar to the challenges of data integration, multi-scale modeling is constrained by limitations in: (i) the schema for the acquisition, representation and measurement of appropriate biological knowledge that can be integrated with quantitative modeling and analysis efforts; (ii) computational and theoretical modeling methods that cross scales and that can be validated; and (iii) computational infrastructure in which factors such as speed, power, data standards and tools for data analysis and visualization do not anticipate the needs of biologists. Cultural and organizational factors also have a role in keeping biological modeling from taking a more prominent role in biology. There is often a lack of adequate multi-disciplinary expertise that is required for developing integrated systems models. There is also a lack of mechanisms to encourage and reward mutually beneficial collaborations between quantitative modelers and biologists. Frequently, opportunities for career advancement and other incentives are not available to computational and theoretical scientists in the biomedical field. Peer review activities for biomedical modeling have large challenges in establishing well-balanced, multidisciplinary review committees as well as review criteria that properly assess the success of model outcomes and validation studies.

The future of modeling in digital biology will necessitate the formation of multidisciplinary collaborations and a level of cooperation that will challenge the research community. The current move towards sharing software and data is expected to provide a significant impetus for the development and implementation of multi-scale modeling. We expect that modeling and simulation of complex biological systems across scales will have an increasingly important role in advancing biomedical and clinical research.

Networked science

Advances in information technology enable biomedical sciences to collaborate more effectively, thereby promoting a richer understanding of fundamental biological phenomena. This technology can also forge a seamless pipeline of scientific information in which the informatics of discovery dovetail neatly into clinical informatics.

Biomedicine is undergoing a transformation in which genomic tools for molecular, structural and functional

analysis are becoming widely available and are being applied to diverse research tasks from modeling biological systems to clinical diagnostics. Linking basic and clinical research through networks will permit scientists to characterize the function of genetic networks in processes of cellular development, health and disease. Such networks will require new interfaces, as well as new translational 'middleware' to merge disparate data sources into usable knowledge. This is indeed a significant challenge. However, the synthesis of insights through collaboration among several biomedical specialties is certain to advance evidence-based medicine and ultimately to improve care.

Today, data required for diagnosis and treatment are acquired in very different ways, across many levels of an organization. These data are stored in various different formats with different standards; moreover these data have different scales with different degrees of resolution and noise. New requirements are arising even as we work toward a truly effective translational medicine. Specifically, those in clinical settings are finding a need for dealing with multi-scale, complex data and to bring data mining, federation and other modern computer networking strategies to electronic medical record keeping. These issues are central to the NIH Roadmap effort to Reengineer the Clinical Research Enterprise (<http://nihroadmap.nih.gov/clinicalresearch/index.asp>) [1].

In the future, digital biology will be comprised of dense networks of people and resources. Advances in instrumentation, collaboration, data query and analysis offer an array of ways to enhance diagnosis, doctor-patient interaction and other aspects of health care delivery. New grid-based methods allow sharing of geographically dispersed heterogeneous data, software algorithms and computing resources (The 2003 Bioengineering Consortium Symposium June 23rd and 24th at NIH focused on team science and the meeting agenda, final report and related publications can be found at <http://www.becon.nih.gov/symposium2003.htm>). The NIH Biomedical Informatics Research Network (BIRN) (NIH National Center for Research Resources Biomedical Information Research Network; www.nbirn.net) is a key example of how grid computing teamed with web-based services can bring informatics-based technologies closer to the scientist's workbench. This initiative fosters large-scale biomedical science collaborations via emerging cyber infrastructure (high speed networks and distributed high-performance computing, as well as the necessary software and data integration capabilities). The new software and hardware technologies are sophisticated and certainly need improvements in areas such as reliability and ease of use. However, there is increasing awareness that technology alone is not the barrier to effective networked science. There is an urgent need for improved interdisciplinary training, effective management principles for team science, proper resolution of government's role in large-scale biomedical data acquisition and analysis efforts and usable solutions to intellectual property issues of data and software sharing and dissemination – in other words, the sociology of networked science is a principal barrier to progress. The Research Teams of the Future component of

the NIH Roadmap is attempting to deal with some of these issues. (<http://nihroadmap.nih.gov/>).

The future of digital biology

The vision of digital biology emerging among scientists today is that of a robust computational environment that supports wide-scale, intense scientific collaboration. The ideal environment for digital biology supports the widest variety of tasks – data analysis, simulation, model building and evaluation – for multiple investigators on a simultaneous basis. It uses models and unique representations of biological data at all levels to link theory with practice. To support the transition of biology to its envisioned future as a more quantitative and a predictive science, the computational environment of digital biology must be reliable, extensible and interoperable in ways not yet realized.

Consider just a few scenarios for how the challenges of data integration, modeling and scientific networks might be addressed. First, universal data standards are unlikely to be adopted widely in the near future, so data capture and translation between multiple sources will demand special attention. Second, biomedical research might be advanced in dramatic ways by robust predictive modeling of complex systems that are validated by iterative interaction between experiment and theory. Third, in addition to new technologies, digital biology will depend on new organizational structures – for example team-based science – as well as partnerships between scientists, government and industry at many levels. A partial future vision of digital biology is illustrated in Box 2: Selected examples of best practice in computational biology. Finally, standards and modular software might be developed to provide an interface for heterogeneous datasets that can be used productively by biologists.

Other aspects of the future of digital biology might include the following:

(i) Semantic interoperability

Coordinated ontologies and terminologies in biology will be established to enable new opportunities for data sharing and heterogeneous data integration. It is likely that many or most research projects that generate data of broad interest will use well-formed terminologies consistent with standards advanced under the Consolidated Health Informatics Initiative (www.whitehouse.gov/omb/egov/gtob/health_informatics.htm). As a part of this trend, open source software development procedures with demonstration projects will proliferate to promote proof of principle and adoption of computational algorithms in the scientific and clinical communities. Such open source development is a requirement for the newly funded National Centers for Biomedical Computing. The National Centers for Biomedical Computing are part of the NIH Roadmap and information about the program and the first round of funded centers is available at <http://www.bisti.nih.gov/ncbc/index.cfm>.

(ii) Data services

Research data infrastructures will be developed to support software tool development and heterogeneous data

Box 2. Selected examples of best practice in computational biology

1. A team of researchers from Case Western Reserve University (Cleveland, Ohio; <http://www.csuohio.edu/mims/index.htm>) is combining computational modeling with physiological experimentation to understand the relationship between metabolism of single human cells and organ and whole body metabolism. This work is yielding computer models of metabolism in liver, heart and brain that promote evidence-based methods for clinical decision support, including diagnosis and treatment [9].

2. An industrial team at United Devices, Inc. (Austin, Texas; <http://ud.com/rescenter/> and http://ud.com/rescenter/files/ds_smallpox.pdf) developed technology for massive computational screening of lead drug compounds for drugs by accessing otherwise unused computer time in a global collaborative network of desktop computers. Recently they reported that this work yielded new compounds against a smallpox protein. This work will bring new drugs into animal and human testing cheaply and quickly, yielding more effective, less expensive drugs (United Devices, Inc. http://www-unix.gridforum.org/7_APM/LSG.htm; www.ud.com/rescenter/files/ds_smallpox.pdf).

3. A team from the University of Connecticut in Storrs, Connecticut (<http://www.cbic.uchc.edu/index.html>) formed the National Resource for Cell Analysis and Modeling, a nationally accessible computational environment for modeling cell functions. This environment speeds the pace of research at the cellular level by permitting researchers to readily put experimental biochemical data in the context of a computational model of a cell to understand how individual biochemical reactions give rise to coordinated functions at the pathway and cellular level [10].

4. A team from Johns Hopkins University (<http://www.bme.jhu.edu/labs/levchenko>) is using Monte Carlo modeling to predict biochemical signaling pathways in heart muscle cells. By using the computer-driven random walk to simulate diffusion of signaling molecules in the cell, it is possible to model cellular behavior in great detail, and thus provide a more detailed view of cell signaling. Cell signaling relates to basic and clinical research [11].

5. A team from Indiana University (<http://www.indiana.edu/~neurosci/sporns.html> and <http://www.indiana.edu/~cortex/robots.html>) is developing an autonomous computational robot with learning capabilities similar to the human brain. This research is aimed at understanding principles of brain function and also at understanding brain function to build automated intelligent systems and robots that can serve human needs [12].

6. A team based at Massachusetts General Hospital/Harvard Medical School is studying malignant brain tumors as self-organizing and adaptive biosystems. Their Tumor Complexity Modeling Project (TCMP) uses methods from various disciplines, such as tumor biology, bioengineering, materials science, mathematical biology, nonlinear physics as well as computational and complex systems science. The immediate aim of TCMP is to develop novel experimental, computational, mathematical and theoretical tumor models. The ultimate goal is to develop virtual treatment planning devices and strategies for malignant brain tumors (<http://btc.mgh.harvard.edu/TumorModeling/>)

integration. Various data repositories will implement software tools to allow format inter-conversions and make data interoperable. Funded projects will routinely make data available. Stakeholders will insist on infrastructures and practices that maintain data for current and future uses. Community-managed databases and data services are probable future developments. These will involve public-private partnerships to improve software development and validation or to accelerate diffusion of useful algorithms. For example, the Lung Image Database Consortium (<http://imaging.cancer.gov/programsandresources/InformationSystems/LIDC>) [2]

gathers standard, annotated images that will be used for imaging research and for a range of research, including the development of tools for clinical decision support ('personalized medicine') or population-based studies of biomedical and scientific phenomena.

(iii) Changed publishing practices

Scientific journals will increasingly use standardized language and document structures in research publications. Many will create companion versions of articles as databases for context-dependent cross-querying of literature. Published reports of experimental data (including negative experimental results) will provide new levels of detail of protocols, so that results are reproducible. Furthermore, articles on technology development should gain greater acceptance in scientific journals to foster multidisciplinary research, such as multi-scale modeling.

(iv) Multi-scale modeling

Biomedical research ultimately requires predictive modeling of complex systems that is experimentally validated. Interdisciplinary projects will look at problems across multiple scales – for example, developing models of ion channel mutations in the heart that can predict the development of disease in the heart. Success in multi-scale modeling depends on the development of novel theoretical and computation approaches to discover new methods for crossing the boundaries between scales. These enabling technologies need to be developed simultaneously with focused data collection at multiple scales for each disease or organ system.

(v) Biomedicine-specific infrastructure

It is realistic to expect that novel solutions adopted in specific biomedical domain problems will be incrementally integrated with prototyped components of the Grid infrastructure. Scientists and administrators will identify and solve performance limitations in grid computing. Provided the investment follows to correct these limitations, the computational infrastructure to support collaborative science will improve and the capacity to manage and share complex, voluminous data and computing resources will be greatly enhanced.

(vi) Support for continuous learning

Education and training are needed to facilitate the evolution of biology toward a large-scale paradigm. Community involvement is required so that the scientific resources made available by large-scale facilities can be used to solve the most problems. Education is needed at two levels: How can large-scale resources and facilities drive scientific progress for an individual investigator? In addition, we need to explore ways to use the tools, data types and the associated dynamic information to gain a higher-level understanding of the problems facing an entire field. Concrete examples are the needs in multi-scale modelling. It is important to train mathematical biologists as well as biological mathematicians.

(vii) Changes in policy and the culture of scientific collaboration

There is a general consensus that regardless of the physical distance between different scientists in the community, frequent interactions and intense data sharing are essential in an era of digital biology. Hence, a wide variety of organizational and policy issues must be addressed, including openness, data dissemination, semantic interoperability, data integrity and validation, protection of human data, operational efficiency, intellectual property and ethics policies, such as conflict of interest [3].

Conclusions

With the advent of genomics and proteomics, bioscientists increasingly acknowledge the ever present need for computers. But few biologists view computers as a cornerstone of biology. Instead most regard computation as a 'black box' and focus on data and analytical outputs, not the machine processes by which they were generated. The catch phrase 'digital biology' points to a fundamentally unique future in which computers will enable scientists to:

Exploit opportunity

To the extent that we approach biology with the proper computer tools and methods, genomics and proteomics will lend themselves to discrete observations and mathematical formulations that were not possible just a few years ago. Our capacity for building robust models of biological processes expands because of the new data and will profoundly extend our questioning and perception.

Avoid problems

Focusing effort on data integration, biological modeling and networked science along the lines outlined above will help to avert a potential crisis owing to the complexity of the biological systems as we are now coming to view them. Without computational support, sciences rooted in genomics and proteomics risk being hindered by the volume and disparate nature of data to be accessed, queried and managed.

Realize a dream

Computation and quantitative methods are at the forefront of life sciences research. The envisaged future would extend human perception in the laboratory, establish the spatial and temporal context of disease at multiple levels, assist scientists with integration of complexity and maintain or create new levels of efficiency in basic discovery processes. With proper multidisciplinary effort, computational and quantitative approaches to biomedical research will produce scientific breakthroughs that lead to

significant health benefits. As revealed through the links in this article, computers have already extended the capacity of physicians to diagnose and anticipate the course of cancers, as well as diseases that affect almost all organs of the body.

Accelerate the pace of scientific advancement

Attending to these issues now is consistent with other recent trends such as intensifying collaboration in biology. It opens the door for biology to draw more deliberately on engineering and other sciences.

Many assert that we are at a crossroads where immediate, decisive action is required to achieve the full potential of digital biology. Whatever specific steps are taken to advance a digital future for biology, scientists should consider the many current challenges in data integration, multi-scale modeling and the networking of science.

Acknowledgements

The authors wish to thank all those who contributed to the discussions that formed the basis for this essay, including Emory Brown, Sydney Brenner, Lisa Brooks, James Cassatt, Mark Ellisman, Robert Fowler, Deb Gracio, Ken Kennedy, Nathan Myhrvold, Larry Smarr, Rick Stevens, Gary Strong, Shankar Subramaniam, Steven Wiley, John Wooley, William Yasnoff and Angela Hight-Walker.

References

- 1 Zerhouni, E. (2003) The NIH Roadmap. *Science* 302, 63; <http://nihroadmap.nih.gov/>
- 2 Armato, S.G., III. *et al.* (2004) Lung Image Database Consortium: Developing a Resource for the Medical Imaging Research Community. *Radiology* 232, 739–748; <http://imaging.cancer.gov/programsandresources/InformationSystems/LIDC/page8>
- 3 Arzberger, P. *et al.* (2004) An International Framework to Promote Access to Data. *Science* 303, 1777–1778
- 4 Goldberger, A.L. *et al.* (2000) PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* 101, 215–220; <http://www.physionet.org/>
- 5 Gibbs, R.A. *et al.* (2003) The International HapMap Project. *Nature* 426, 789–796; <http://www.hapmap.org/>
- 6 Atkins, D.E. *et al.* (2003) Revolutionizing Science and Engineering through Cyber infrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyber infrastructure National Science Foundation, Arlington, VA
- 7 Zerhouni, E. (2003) The NIH Roadmap *Science* 302, 63; <http://nihroadmap.nih.gov/>
- 8 Interagency Opportunities in Multi-Scale Modeling in Biomedical, Biological, and Behavioral Systems. Program Solicitation (2004) NSF 04-607
- 9 Salem, J.E. *et al.* (2004) Step and ramp induction of myocardial ischemia: Comparison of *in vivo* and *in silico* results. *J. Physiol. and Pharm.* 55, 519–536; <http://www.csuohio.edu/mims/index.htm>
- 10 C.P. Fall *et al.* eds (2004) Computational Cell Biology, Springer-Verlag. <http://www.cbit.uchc.edu/index.html>
- 11 Murrell, M.P. *et al.* (2004) The systems biology of glycosylation. *Chem. Bio. Chem.* 5, 1334–1347; <http://www.bme.jhu.edu/labs/levchenko/>
- 12 Sporns, O. and Kotter, R. Motifs in Brain Networks. *PLoS Biology*, (in press); <http://www.indiana.edu/~neurosci/sporns.html>; <http://www.indiana.edu/~cortex/robots.html>